PROCEEDINGS *of the* SECOND BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY

Held at the Statistical Laboratory
Department of Mathematics
University of California
July 31–August 12, 1950

EDITED BY JERZY NEYMAN

UNIVERSITY OF CALIFORNIA PRESS
BERKELEY AND LOS ANGELES
1951

UNIVERSITY OF CALIFORNIA PRESS
BERKELEY AND LOS ANGELES
CALIFORNIA

CAMBRIDGE UNIVERSITY PRESS
LONDON, ENGLAND

PRINTED IN THE UNITED STATES OF AMERICA

EXPERIMENTAL CORRELOGRAM ANALYSES OF ARTIFICIAL TIME SERIES (WITH SPECIAL REFERENCE TO ANALYSES OF OCEANOGRAPHIC DATA)

H. R. SEIWELL
WOODS HOLE OCEANOGRAPHIC INSTITUTION

CONTENTS

1. Introduction

When dealing with data as they come from nature, the applied statistician is frequently faced with the prospect of selecting and modifying theoretical mathematical methods to fit the special conditions imposed by the data themselves. Since theoretical mathematical tools are developed for specific situations they may become unsafe for applications outside the framework for which they were originally intended. Observations on natural phenomena cannot usually be controlled in the sense of a controlled laboratory experiment, and for obvious reasons, they are finite in scope. Hence, some form of experimental modification of theoretical formulae is usually essential if the maximum correct information is to be obtained from the data.

Data dealing with time variations of natural phenomena usually leave much to be desired in that the lengths of observed series may be woefully short, they are masked by various degrees of error and are sometimes sporadic and disconnected. Furthermore, the dynamics of the generating mechanisms are frequently not known, so that *a priori* considerations are purely speculative. A further complication concerns the length of the time series to be utilized for computation. The individual members of the series being dependent, one on the other, so that regardless how many are chosen, within practical limits, the number becomes small in light of the requirements imposed by random sampling. A further point of

This paper is contribution No. 531 from the Woods Hole Oceanographic Institution and U.S. Navy Office of Naval Research Contract No. N6onr 277 Task Orders 1 and 8. At the time of his death on March 7, 1951, the author was a John Simon Guggenheim Memorial Fellow, 1950–1951.

significance is that the observed sequence of events may be stationary only for brief intervals and, hence, imposes a limit to the length of series entering into the analysis.

Experience has indicated the superiority of the correlogram method in certain analyses of geophysical time series. However, it is new and practical experimentation is required to interpret the results revealed by its application to natural data. Hence, at a risk of being considered not too fashionable, an experimental study of correlogram analyses of several artificially generated series has been undertaken. The following is a descriptive account of the first results of these researches on four mathematical models. The information obtained has been very useful in the interpretation of correlogram analyses of natural data with which we are concerned. Although experience with many analyses has revealed the power of the correlogram method, general conclusions have at this time been avoided. It is to be hoped that the following results will stimulate additional research.

2. Computational procedure

Both hand and machine procedures for computation of autocorrelation coefficients, as undertaken in this laboratory, have been previously described [7].

The large amount of computation required for the investigation was under the supervision of Mr. Thomas C. Duke and Mrs. Mary M. Hunt.

3. Experimental model I

3.1. *Pattern of analyses.* Model I is an analysis of the function

$$y = \sin \frac{2\pi x}{10} + \epsilon \tag{1}$$

where ϵ is an assumed rectangularly distributed random variable.

The object of the study was to obtain general information on the correlogram of the function and to observe the effect of a changing stochastic variable on statistical properties of both the basic data and the correlogram. Variations in the function studied were:

$$\text{Series A} = \epsilon$$
$$\text{Series B} = 5\epsilon$$
$$\text{Series C} = 8\epsilon$$
$$\text{Series D} = 10\epsilon .$$

Each series, composed of a sine wave plus added random numbers, may be likened to a message containing a basic signal masked by a degree of noise. A fundamental problem is to isolate the signal from the message and to discover as much as possible about its physical properties. Although the models have no particular physical significance, they may approximate a type of function underlying natural time series. In the special case of Oceanography, it appears reasonable that at times the sea surface roughness pattern may be represented by a single cyclical component (sea swell) plus random disturbances (local sea) [5]. However, regardless of the exact nature of the phenomenon, it is significant in the beginning to dem-

onstrate that for series of the type represented by model I, conclusions regarding the physical properties of the phenomena require special transformations of the basic data.

3.2. *Properties of the basic data.* Statistical properties of the basic data (figure 1) tabulated in table I are based on a minimum of twenty-five complete cycles (200 units). The computed means and variances for the series do not differ significantly from theoretical values and the samples closely approximate properties of the parent.

The presence of the random disturbances masks the underlying rhythmic cycle

Series A

Series B

Series C

Series D

FIGURE 1

Graphs of model I data

and it will not be recognized by the distances between successive peaks or successive upcrosses. Thus, for series A, where the cosine variance is 86 percent of the total variance, the mean distance between peaks is not 10, but 4.86 units. This distance between peaks becomes less as the magnitude of randomness is increased, and at the five fold mark, its mean value is but little more than the theoretical value of 3 units, characteristic of a random sequence.

Distances between upcrosses indicate essentially the same situation; for series A, the mean distance of 8.76 units rapidly diminishes with increased randomness to the theoretical value of 4 units, characteristic of a random series.

Hence, it is apparent that when the underlying rhythmic component of a simple time series is disturbed by a random influence (noise), the properties of the component (or even its presence) are not revealed by the method of measuring distances between peaks or upcrosses in the basic data. This information is not new. It has been discussed by other investigators, and in particular, M. G. Kendall [2] has shown its unreliability for several well known economic series. The argument that selective methods may be used for identifying peaks appears equally unsound unless the effect of the randomness can be first eliminated.

TABLE I

SELECTED STATISTICAL PROPERTIES OF MODEL I

PROPERTY	SYMBOL	COMPUTED VALUES				THEORETICAL VALUES			
		Series A	Series B	Series C	Series D	Series A	Series B	Series C	Series D
Distribution of ϵ		R(0, 1)	R(0, 5)	R(0, 8)	R(0, 10)	.....	.....	.....	.....
Function mean	M	0.516	2.539	4.001	5.068	0.500	2.500	4.000	5.000
Average deviation	AD	0.667	1.330	2.120	2.542	.....	.....	.....	.....
Standard deviation	σ	0.773	1.579	2.523	2.914	0.764	1.607	2.415	2.972
Deviation ratio	AD/σ	0.863	0.842	0.840	0.872	.....	.....	.....	.....
Mean dist. peaks		4.86	3.03	3.01	2.94	9.13	6.14	4.80	3.56
Mean dist. upcrosses		8.76	3.90	4.08	4.06	9.99	9.65	8.62	5.24
Total variance	σ_x^2	0.598	2.492	6.366	8.727	0.583	2.582	5.831	8.833
Cosine variance	$\sigma_{\cos}^2$	.....	.....	.....	.....	0.500	0.500	0.500	0.500
ϵ variance	σ_ϵ^2	0.092	2.1212	5.856	8.485	0.083	2.082	5.331	8.333
Period cosine	T	10.0	10.0	10.0	.....	10.0	10.0	10.0	10.0
Amplitude cosine	C	0.991	0.854	1.128	0.778	1.00	1.00	1.00	1.00
Terminal amplitude	r_T	0.89	0.19	0.12	0.0	0.857	0.194	0.086	0.057

To carry the above a little further, theoretical distances between peaks and upcrosses of the four series were computed by the method of M. G. Kendall, which is applicable to a simple linear autoregressive series of the type

$$u_{t+2} + au_{t+1} + bu_t = \epsilon_{t+2} \tag{2}$$

where ϵ is a random variable, and u_t normally distributed. This method and procedure is discussed in detail in the section on autoregression, to which it is applicable. The results obtained from its application to model I are tabulated in table I. They do not, as expected, throw any particular light on the subject and are presented only for future reference. However, it is to be noted that for series A, the

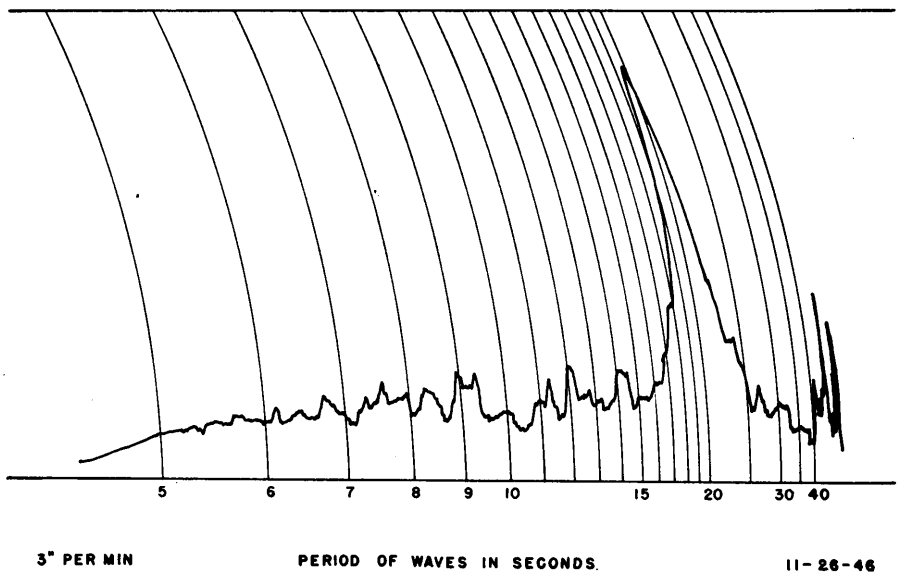

FIGURE 2

Periodogram model I, series A; period scale 2 × units

computation, based on the first two autocorrelations, gives correct period for distance between upcrosses.

The masking effect of the random component on amplitudes of the trigonometric is also revealed by data in table I, where the sine amplitude C computed by least squares for a ten unit period in the basic data, attained 99 percent of its value for series A, and diminished to 78 percent for series D.

Other statistics of the basic data are tabulated in table I for reference purposes. The ratio of the average deviation to the standard deviation is significant in that it lies between 0.84 and 0.87, a value fairly close to the theoretical AD/σ ratio for cosines. This property is discussed in a later section.

3.3. *Periodogram analyses.* Classical periodogram analysis of natural time series, for which there is no *a priori* knowledge of periodicity, has fallen into more or less disrepute. The results of periodogram analyses performed on each of the four series (figures 2 to 5) by means of a mechanical periodogram analyzer [4], are

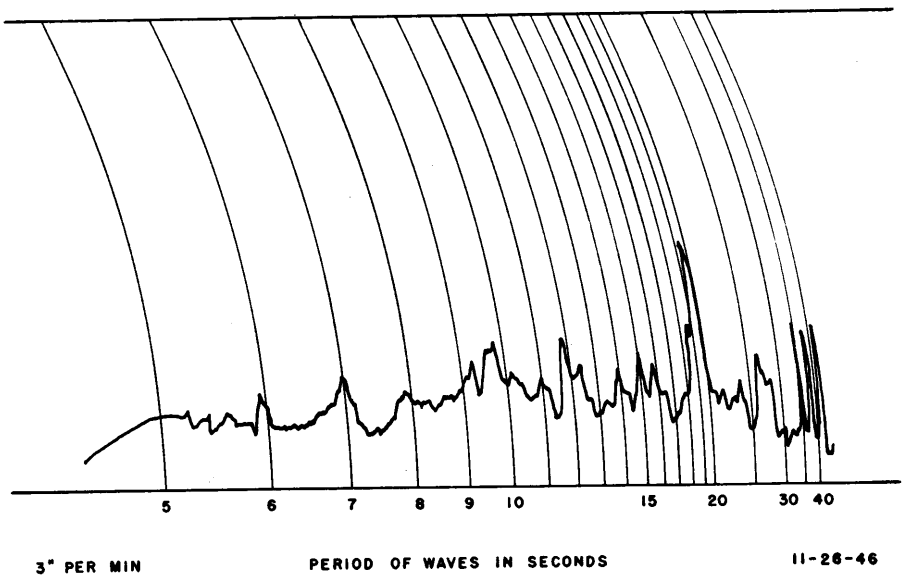

FIGURE 3

Periodogram model I, series B; period scale 2 × units

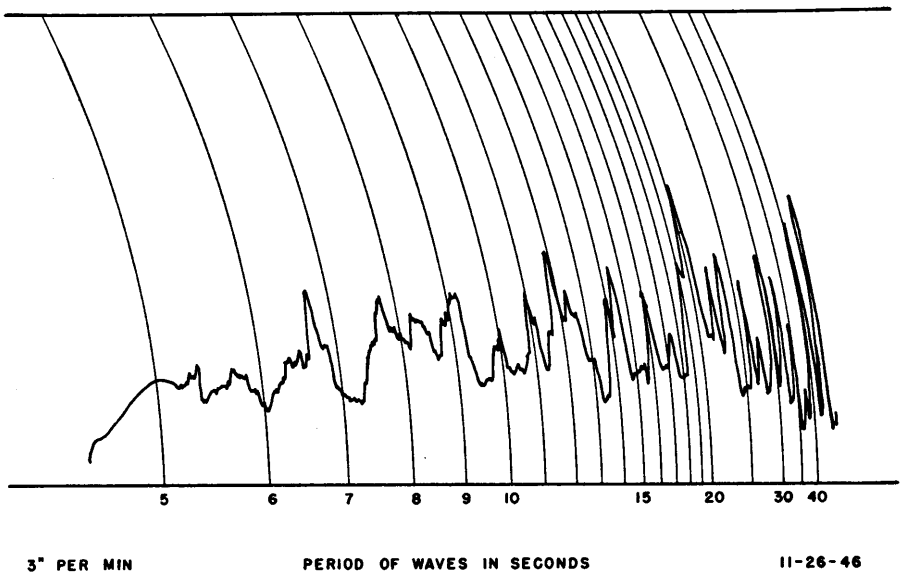

FIGURE 4

Periodogram model I, series C; period scale 2 × units

presented to illustrate the danger of misinterpreting fundamental properties of natural time series.

In the case of series A (variance of random component 14 percent of the total variance) the ten unit peak is clearly defined in its periodogram and provided minor peaks are ignored, the analysis could be correctly interpreted. However, as the magnitude of the random component increases, as in series B (variance of random component 81 percent of total or about five times the cosine variance) the relative magnitudes of the spurious peaks increase, and although the ten unit peak is still superior, there is danger of interpreting the periodogram as representing an interference pattern comprised of other frequencies, particularly those between 5 and 30 units.

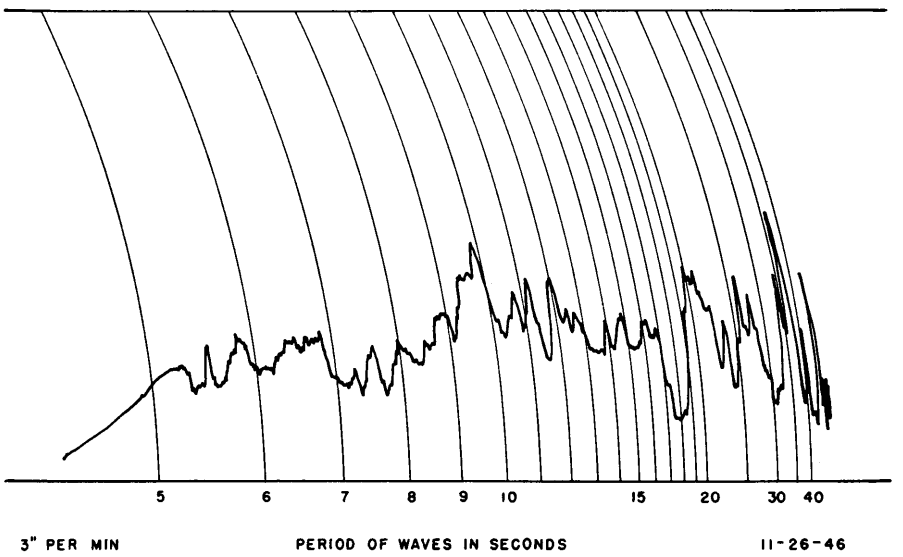

FIGURE 5

Periodogram model I, series D; period scale 2 × units

As the magnitude of the random component increased still further, the periodogram became more complex, and for series C (variance of random component approximately 12 times the cosine variance) the dozen or more principal peaks could be interpreted as indicating an equal number of harmonic components in the basic data.

In the final case of series D (variance of random component approximately 17 times the cosine variance), the ten unit peak is completely masked and unrecognizable in its periodogram.

Periodograms similar to these frequently result from the analyses of natural data for which there is no *a priori* knowledge of periodicity and have been interpreted as providing clues to physical properties of the data [1]. The subject is not new and various investigators have discussed the fallacy of conclusions from periodogram analyses of economic and natural time series [2], [8]. Our experiments indicate that periodogram analyses of finite data may possibly be interpreted with

a degree of reliability only when the random component is absent or of very low magnitude, and only then by the point location of the most prominent peak on the frequency scale. Its breadth is not sufficient evidence of a train of waves in the basic data.

The above analyses were undertaken on finite series comprised of 215 units. Although some deleterious effects of periodogram analyses are eliminated by using larger and larger amounts of data, a situation is rapidly approached where the computational labor becomes prohibitive.

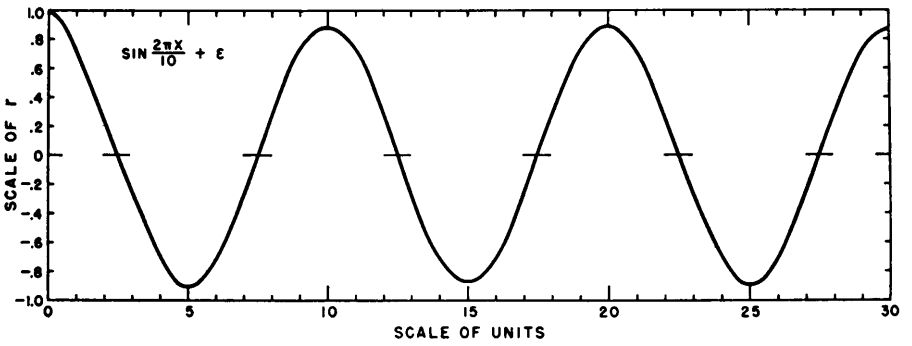

FIGURE 6
Correlogram model I, series A

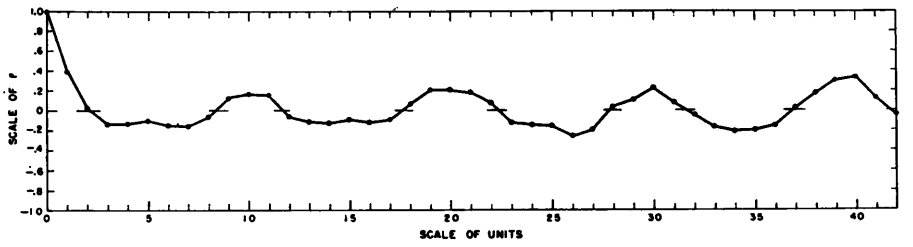

FIGURE 7
Correlogram model I, series B

3.4. *Correlogram analyses.* (a) *Basic data.* In these analyses we follow a pattern developed for ocean wave observations. After preliminary study, the basic data are subjected to autocorrelation analyses, the correlogram is drawn and then evaluated. The particular procedure applied to model I is applicable to natural data only in the special case where its correlogram clearly reveals a single rhythmic component and when it damps to a terminal amplitude. When this is the case, the amplitude, associated with the frequency determined by autocorrelation, is computed and the residue, remaining after its subtraction from the basic data, is again subjected to autocorrelation analyses.

The correlograms of the four series (figures 6 to 9) when viewed together clearly demonstrate the power of autocorrelation analysis as a means of revealing the period of the rhythmic component in the basic data, provided it is not completely

masked by the random component. Thus, for series A and B, the ten unit period is clearly defined by the correlogram; in series C, it may be estimated with reasonable accuracy; but, in series D, the random variable is too large to permit period identification. When natural data are being investigated, periods revealed in this fashion are utilized for determining the amplitude of the rhythmic component.

The correlograms of series A, B and C, are characteristic of those for a single

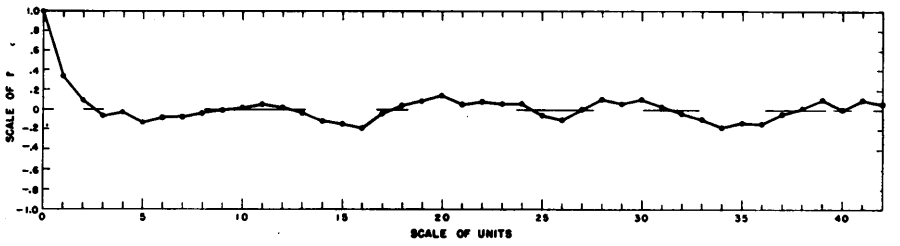

FIGURE 8
Correlogram model I, series C

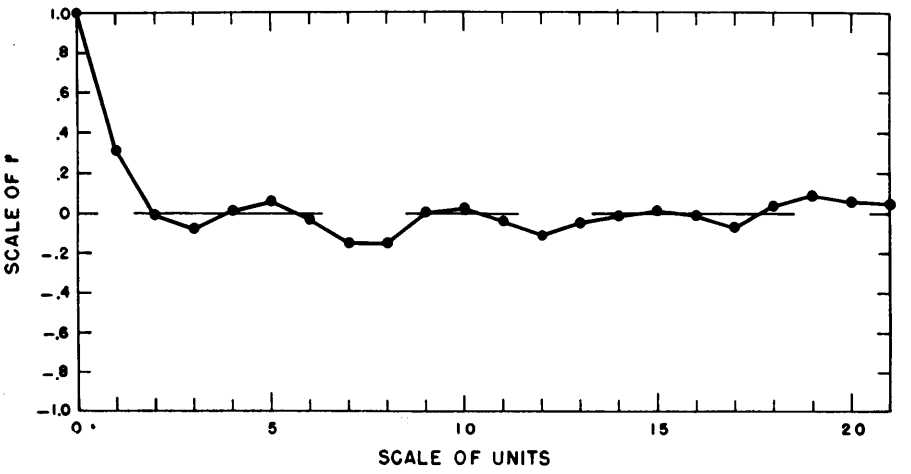

FIGURE 9
Correlogram model I, series D

trigonometric with added random noise. They damp to terminal amplitudes and remain constant. This, together with the uniform correlogram period of ten units, provides information on physical properties of unknown series and is similar to a type of correlogram which has been obtained for sea surface wave heights [5], [6]. The terminal amplitudes observed for the correlograms agree well with their theoretical values (table I) and demonstrate the reliability of the autocorrelation computation on finite amounts of data for this type of function when not completely masked by random noise.

The theoretical terminal amplitude r_T of a sine with added random component is

$$r_T = \frac{\text{variance cosine}}{\text{total variance}}. \tag{3}$$

It expresses the percent reduction due to the cyclical component and is an important property of the data.

(b) *Residual data.* The correlogram for the residuals of series A (figure 10) is similar to that for a series of random (or nearly so) numbers. It demonstrates that practically complete removal of the cosine from the basic data is possible under the circumstances outlined.

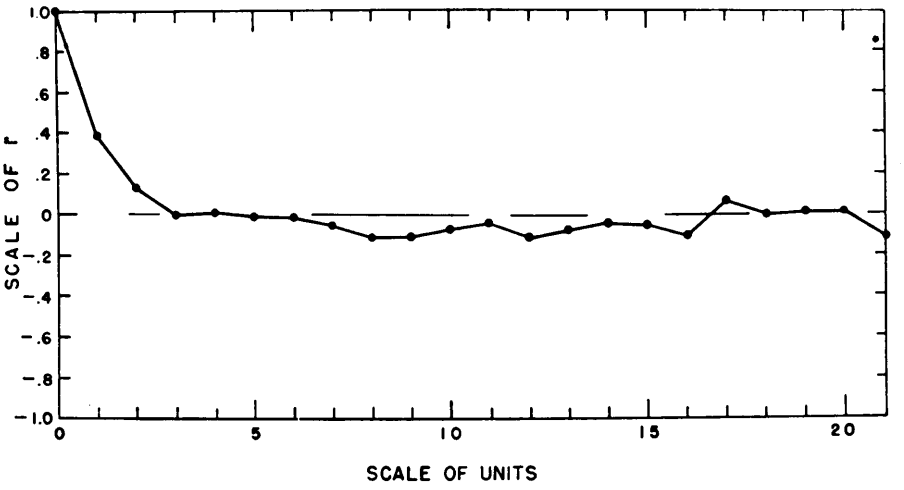

FIGURE 10

Correlogram of residuals model I, series A

4. Experimental model II

4.1. *Pattern of analyses.* Model II is concerned with analyses of combinations of a sum of sines:

$$y = a_1 \sin \frac{2\pi}{T_1} k + a_2 \sin \frac{2\pi}{T_2} k + a_3 \sin \frac{2\pi}{T_3} k . \tag{4}$$

It is known that the correlogram of a series of several harmonic terms will have a sinusoidal form which does not damp to zero. However, the form of the correlogram varies according to the combination of cyclical components making up the original series. It is important that the properties of correlograms of such combinations be examined for distinguishing characteristics to provide a general guide in correlogram analyses of natural series.

In this section we are concerned only with the forms of the correlograms drawn in each case from sixty theoretical autocorrelation coefficients r_k in one unit steps, and computed according to the general equation:

$$r_k = \frac{a_1^2 \cos \frac{2\pi}{T_1} k + a_2^2 \cos \frac{2\pi}{T_2} k + a_3^2 \cos \frac{2\pi}{T_3} k}{a_1^2 + a_2^2 + a_3^2} . \tag{5}$$

The various combinations of harmonic terms considered in this section are as follows. Key characteristics of their correlograms (illustrated in figure 11) are tabulated in tables II and III.

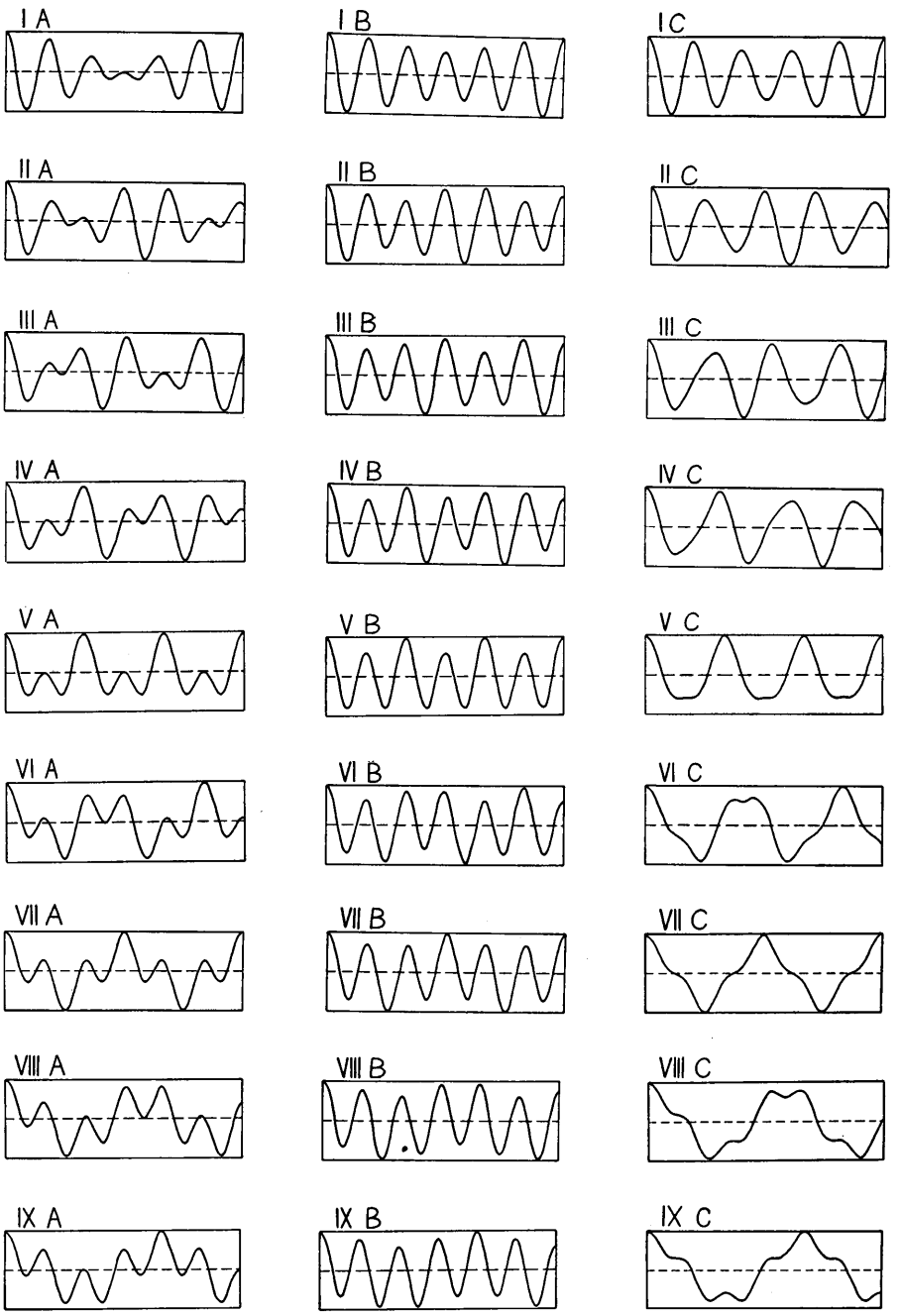

FIGURE 11

Correlograms of combinations of sums of two sines of model II. See table II for period and amplitude combinations. In each correlogram, abscissa is scale of units from 0 to 60 and ordinate is scale of r from -1.0 to $+1.0$. Left column—series I–A to IX–A; middle column—series I–B to IX–B; right column—series I–C to IX–C.

Series A. Nine combinations of two sines

$$a_1 = a_2 ,$$
$$T_1 = 10 , \qquad T_2 = 12 \text{ to } 40 .$$

Series B. Nine combinations of two sines

$$a_1 = 1 , \qquad a_2 = 0.5 ,$$
$$T_1 = 10 , \qquad T_2 = 12 \text{ to } 40 .$$

Series C. Nine combinations of two sines

$$a_1 = 1 , \qquad a_2 = 2 ,$$
$$T_1 = 10 , \qquad T_2 = 12 \text{ to } 40 .$$

Series D. Seven combinations of three sines

$$a_1, a_2, a_3 = \text{combinations of 1 and 2 ,}$$
$$T_1 = 10 , \quad T_2 = 12 , \quad T_3 = 14 .$$

Series E. Seven combinations of three sines

$$a_1, a_2, a_3 = \text{combinations of 1 and 2 ,}$$
$$T_1 = 10, \quad T_2 = 12 , \quad T_3 = 40 .$$

Series F. Three combinations of three sines

$$a_1, a_2, a_3 = \text{combinations of 1 and 2 ,}$$
$$T_1 = 10 , \quad T_2 = 20 , \quad T_3 = 30 .$$

4.2. *Summary of correlogram characteristics for series composed of more than one cyclical component.* The correlograms of series consisting of two sines with various amplitude and period combinations have unequal distances between cycles, they do not damp to zero, but show alternating damping and growth. Series of this type generally are recognizable within the first few correlogram cycles, as distinguished from correlograms of series having a single cyclical component. Unequal distances between correlogram peaks always occurred within the first four cycles, the usual difference of 1 to 3 units occasionally attained 23 units. As the periods of the two sine combinations diverge, distances between successive peaks of the correlogram tend to become more uniform but the alternate damping and growing increases. In the case where the longer period wave has the larger amplitude, the correlogram is less symmetrical, and damping is consistent through 1 to 3 cycles only. It is the irregular nature of the damping and growth and the unequal distances between peaks, usually occurring within the first 3 to 4 cycles, and the occurrence of ripples, which permit identification of the correlogram as representing a basic series containing more than one trigonometric.

Correlograms of series composed of three sines appear to be, in general, more symmetrical than the above; although rippling increases and asymmetry increases when the longer period sine has the larger amplitude. Damping is consistent only through the first 1 to 4 cycles and distances between individual cycles range up to 16 units within the first 2 to 5 cycles. Differences in distances within the first 2 cycles are generally of the order of 0.5 to 3.5 units. In these cases, as in the above, the nature of the damping and of the differences in distances between peaks and

the occurrence of ripples permit some discrimination of the correlogram within the first 3 or 4 cycles.

In practice correlograms of natural time series which indicate the presence of more than one cyclical component are subjected to a Fourier transform of their autocorrelation coefficients. In this manner it is frequently possible to identify the frequencies of the components and evaluate their other properties.

TABLE II

PROPERTIES OF CORRELOGRAMS OF TWO SINES: $a_1 \sin \frac{2\pi}{T_1} k + a_2 \sin \frac{2\pi}{T_2} k$

Series A: $a_1 = a_2$; Series B: $a_1 = 1$, $a_2 = 0.5$; Series C: $a_1 = 1$, $a_2 = 2$

Curve Designation	Column: 1		2	3	4	5	6	7	8	9	10	11
	Unit Periods		Unit Distances between Peaks for Cycles:						Max. Difference	Min. No. Cycles	First Two Cycle Difference	Consistent Damping to Cycle
	T_1	T_2	1	2	3	4	5	6				
I A	10	12	11	10.5	8.5	8.5	10.5	11	2.5	3	0.5	3
II A	10	14	11	8	10.5	11	10.5	8	3	2	3	2
III A	10	16	11	8	11.5	9.5	9.5	....	3.5	3	3	1
IV A	10	18	10.5	9.0	11.5	8	12	8	3.5	4	1.5	1
V A	10	20	10	10	10	10	10	10	0	...	0	1
VI A	10	25	9.5	11	9	11	9.5	9.5	2	3	1.5	1
VII A	10	30	9.5	11	9.5	9.5	11	9.5	1.5	2	1.5	1
VIII A	10	35	9.5	11	10	9	10	10.5	1.5	2	1.5	1
IX A	10	40	9.5	10.5	10.5	9.5	9.5	10.5	1	2	1	1
I B	10	12	10.5	10	9.5	9.5	10	10.5	1	3	0.5	3
II B	10	14	10.5	9.5	9.5	10.5	10	9.5	1	2	2	2
III B	10	16	10.5	9.5	10.5	10	10	10	1.5	2	1.5	1
IV B	10	18	10	9.5	10.5	9.5	10.5	9	1	3	0.5	1
V B	10	20	10	10	10	10	10	10	0	...	0	1
VI B	10	25	10	10	10	10	10	10	0	...	0	1
VII B	10	30	10	10	10	10	10	10	0	...	0	1
VIII B	10	35	10	10	10	10	10	10	0	...	0	1
IX B	10	40	10	10	10	10	10	10	0	...	0	1
I C	10	12	11.5	12	12.5	12	11.5	....	1	3	0.5	3
II C	10	14	13.5	15	13	15	....	....	2	3	1.5	1
III C	10	16	17	14	17.5	....	....	....	3.5	3	3	1
IV C	10	18	19	19	14	....	....	....	5	3	0	1
V C	10	20	20	20	20	....	....	....	0	...	0	0
VI C	10	25	22	6	22	....	....	....	16	2	16	1
VII C	10	30	30	30	....	....	....	....	0	...	0	0
VIII C	10	35	31	8	....	....	....	....	23	2	23	1
IX C	10	40	20	20	....	....	....	....	0	...	0	1

Tables II and III list the data discussed above. The first column tabulates the periods of the two sine combinations (table II) and amplitudes (table III) of the three sine combinations. Columns 2 to 7 tabulate the unit distances between successive correlogram peaks for six cycles and column 8 gives the maximum difference in unit distance between correlogram peaks. Then column 9 lists the minimum number of correlogram cycles in which the column 8 range occurred and in column 10 appears the differences in the first two correlogram peak distances. Finally, the last column tabulates the number of peaks (cycles) over which the correlogram consistently damped before growth.

TABLE III

PROPERTIES OF CORRELOGRAMS OF THREE SINES: $a_1 \sin \frac{2\pi}{T_1} k + a_2 \sin \frac{2\pi}{T_2} k + a_3 \sin \frac{2\pi}{T_3} k$

Top group: $T_1 = 10$, $T_2 = 12$, $T_3 = 14$; *Center group:* $T_1 = 10$, $T_2 = 12$, $T_3 = 40$
Bottom group: $T_1 = 10$, $T_2 = 20$, $T_3 = 30$

Column: 1			2	3	4	5	6	7	8	9	10	11
Amplitudes			Unit Distances between Peaks for Cycle:						Max. Difference	Min. No. Cycles	First Two Cycle Difference	Consistent Damping to Cycle
a_1	a_2	a_3	1	2	3	4	5	6				
1	1	1	11.5	10.5	6.5	11.0	9.5	10.0	5	3	1	2
1	1	2	13	15	13	16.5	....	....	3.5	5	2	2
1	2	2	12.5	13.5	13	9	10.5	....	4.5	4	1	4
1	2	1	11.5	12.5	13	11.5	11	....	2	5	1	3
2	1	1	10.5	9 5	9.5	10.5	9.5	9.0	1	2	1	2
2	1	2	11.5	8	10	11	10	8.5	3.5	2	3.5	2
2	2	1	11	10.5	8.0	9.0	10.5	10.5	3	3	0.5	3
1	1	2	10	13	16	9	12	....	7	4	3	1
1	1	1	10.5	11	10.5	6.5	10.5	11	4.5	4	0.5	1
2	1	1	10.5	10	9.5	9.5	10	10.5	1	3	0.5	1
1	2	2	11	13	13	11	12	....	2	2	2	1
1	2	1	11.5	12	13	12	11.5	....	1.5	3	0.5	1
2	1	2	10	10.5	10	9	10	10.5	1.5	4	0.5	1
2	2	1	10.5	11	9	8	10.5	11	3	4	0.5	3
1	1	1	9.5	12	8.5	9.5	11	9.5	3.5	3	2.5	0
1	1	2	23	7	7.5	22.5	....	....	16	2	16	1
2	1	2	9.5	11	9.5	9.5	11	9.5	1.5	2	1.5	1

5. Experimental model III

5.1. *The autoregressive scheme.* Knowledge of the autoregressive scheme is due primarily to the work of Maurice Kendall [2], [3]. We consider a series defined by the difference equation:

$$u_{t+2} = -a\, u_{t+1} - b\, u_t + \epsilon_{t+2}\,, \tag{6}$$

which may be regarded as expressing the regression of u_{t+2} on u_{t+1} and u_t. The term ϵ_{t+2} being a residual error.

The theoretical correlogram for a series generated by this difference equation will be damped and is given by

$$r_k = \frac{p^k \sin(k\theta + \psi)}{\sin \psi} \tag{7}$$

where the damping factor

$$p = \sqrt{b} \tag{8}$$

and

$$\cos \theta = -\frac{a}{2\sqrt{b}}, \tag{9}$$

$$\tan \psi = \frac{1+b}{1-b} \tan \theta\,. \tag{10}$$

The autoregressive period is:

$$\frac{2\pi}{\theta}. \tag{11}$$

Kendall points out that typical series of this kind have no "periods" in the strict sense. The lengths from peak to peak or from upcross to upcross vary and experiments indicate that the distributions are unimodal having central values somewhere near the mean distances. Hence, this central value of the distribution is considered the "period" of an autoregressive series.

The first two theoretical autocorrelations $r_0 = 1$ in terms of the coefficients are:

$$r_1 = \frac{-a}{1+b}, \tag{12}$$

$$r = \frac{a^2 - b(1+b)}{1+b}. \tag{13}$$

Also,

$$a = -\frac{r_1(1-r_2)}{1-r_1^2}, \tag{14}$$

$$b = -1 + \frac{1-r_2}{1-r_1}. \tag{15}$$

The estimate of the coefficients is quite sensitive to superimposed errors and it is unsafe to estimate the autoregressive period without reference to the correlogram. In our practical wave record analyses we base estimates of the autoregressive periods on values of the autocorrelations as well as on locations of the first valley and first peak of the correlogram.

The theoretical total variance of the autoregressive series is:

$$\sigma_y^2 = \frac{\sigma_\epsilon^2}{1 + a r_1 + b r_2} \tag{16}$$

where σ_ϵ^2 is the random variance.

In the practical applications of the autoregressive scheme to natural time series, certain properties of the basic data, as the mean distances between peaks and upcrosses, are important to identify properties of the series. For the normal series, the theoretical mean distance between peaks is given by:

$$\text{M.D. peaks} = \frac{2\pi}{\arccos \tau} \tag{17}$$

where

$$\tau = \frac{-1 + 2r_1 - r_2}{2(1-r_1)} \tag{18}$$

or

$$\tau = \frac{b^2 - (1+a)^2}{2(1+a+b)} = \frac{b-a-1}{2}. \tag{19}$$

The theoretical mean distance between upcrosses is

$$\text{M.D. upcrosses} = \frac{2}{\arccos r_1}. \tag{20}$$

Thus, for a random series the theoretical mean distance between peaks is 3 and that between upcrosses is 4.

5.2. *Pattern of analyses.* This section is concerned with an experimental correlogram analysis of eight autoregressive series of the type

$$x_{n+2} = -ax_{n+1} - bx_n + \epsilon_{n+2} . \tag{21}$$

Computations of autocorrelations of each series were based on a minimum of 180 to 500 terms. Duplicate computations were carried out on one series (G) using 180 and 500 terms, respectively. Two of the series (I and J) were identical with the exception of the stochastic variable being rectangularly distributed between -1 and 1 in one case and normally distributed between 0 and 1 in the other. Principal variations within the model, each indicated by a series letter, were produced by changing the damping coefficient (b) and holding the autoregressive period constant at 10 units. In table IV are tabulated the coefficients and the data pertinent to describe the various series. Properties of the series are tabulated in table V.

TABLE IV

Series	Constants a	Constants b	Period	Dist. ϵ	No. Terms
B	−1.62	1.00	10	$R(-2, 2)$	215
F	−1.54	0.90	10	$R(-3, 3)$	215
G	−1.45	0.80	10	$R(-1, 1)$	180
G-1	−1.45	0.80	10	$R(-1, 1)$	500
I	−1.14	0.50	10	$R(-1, 1)$	180
J	−1.14	0.50	10	$N(0, 1)$	180
K	−0.89	0.30	10	$N(0, 1)$	180
L	−1.35	0.70	10	$N(0, 1)$	500
M	−1.25	0.60	10	$N(0, 1)$	400

5.3. *The mean distance between peaks and upcrosses.* From equations of the autoregressive period (11), the mean distances between peaks (17) and between upcrosses (20) in the basic data, we have as the ratio of mean distance between peaks T_p and autoregressive period T

$$\frac{T_p}{T} = \frac{1}{T} \arccos \tfrac{1}{2} \left(b + 2\sqrt{b} \cos \frac{2\pi}{T} - 1 \right) \tag{22}$$

and as the ratio of mean distance between upcrosses T_u and autoregressive period T

$$\frac{T_u}{T} = \frac{2\pi}{T \arccos \left[\dfrac{2\sqrt{b}}{1+b} \cos \dfrac{2\pi}{T} \right]} . \tag{23}$$

The theoretical T_p/T and T_u/T ratios for all possible values of the coefficient, b, and the frequency range encountered in our research are illustrated by figures 12 and 13.

It is apparent that only in certain cases will either T_p or T_u alone, provide a reasonable estimate of the autoregressive period. On the other hand, the T_p/T and T_u/T ratios may assist in discovering facts about unknown natural series,

TABLE V

PROPERTIES OF ARTIFICIALLY GENERATED AUTOREGRESSIVE SERIES (SEE TEXT)

Series	B	F	G	G–1	I	J	K	L	M
Function mean M	−0.079	0.046	−0.077	−0.003	−0.068	0.035	0.12	0.100	0.015
Average deviation AD	7.794	6.316	1.120	1.196	0.717	1.314	1.116	1.836	1.660
Standard deviation σ	9.300	7.747	1.374	1.503	0.882	1.685	1.377	2.264	2.063
Deviation ratio AD/σ	0.84	0.81	0.81	0.80	0.81	0.78	0.81	0.81	0.81
Mean dist. peaks (computed)	10.02	8.29	6.56	6.68	4.54	4.85	3.93	6.12	5.33
Mean dist. peaks (theoretical)	10.03	8.19	7.02	7.02	5.05	5.05	4.26	6.17	5.55
Mean dist. upcrosses (computed)	9.99	9.90	9.42	9.56	8.07	7.20	7.28	9.20	9.82
Mean dist. upcrosses (theoretical)	10.03	10.04	9.91	9.91	8.88	8.88	7.69	9.62	9.32
Total variance (computed) σ_y^2	86.49	60.01	1.89	2.26	0.78	2.84	1.90	5.12	4.26
ϵ variance (computed) σ_ϵ^2	1.35	2.79	0.31	0.32	0.31	1.03	0.97	1.01	1.03
r_1 (computed)	0.77	0.80	0.79	0.71	0.73	0.74	0.73	0.76	0.69
r_1 (theoretical)	0.809	0.808	0.804	0.804	0.763	0.763	0.682	0.796	0.783
r_2 (computed)	0.28	0.30	0.30	0.30	0.26	0.36	0.29	0.20	0.22
r_2 (theoretical)	0.309	0.340	0.363	0.363	0.373	0.373	0.305	0.377	0.382
Total variance (theoretical)	.....	48.619	2.675	2.675	1.0524	3.158	2.064	5.376	4.000
ϵ variance (theoretical)	1.333	3.000	0.3333	0.3333	0.3333	1.000	1.000	1.000	1.000

particularly when they are used in conjunction with another statistical property.

The mean distances between successive peaks and successive upcrosses scaled from curves of the basic data, and those computed from equations 17 and 20 (theoretical) for the nine artificial autoregressive series are tabulated in table V. The autoregressive period $2\pi/\theta$ was identical for all series, and the results together with figures 12 and 13 permit the following conclusions.

1. The mean distances between upcrosses are greater than those between peaks

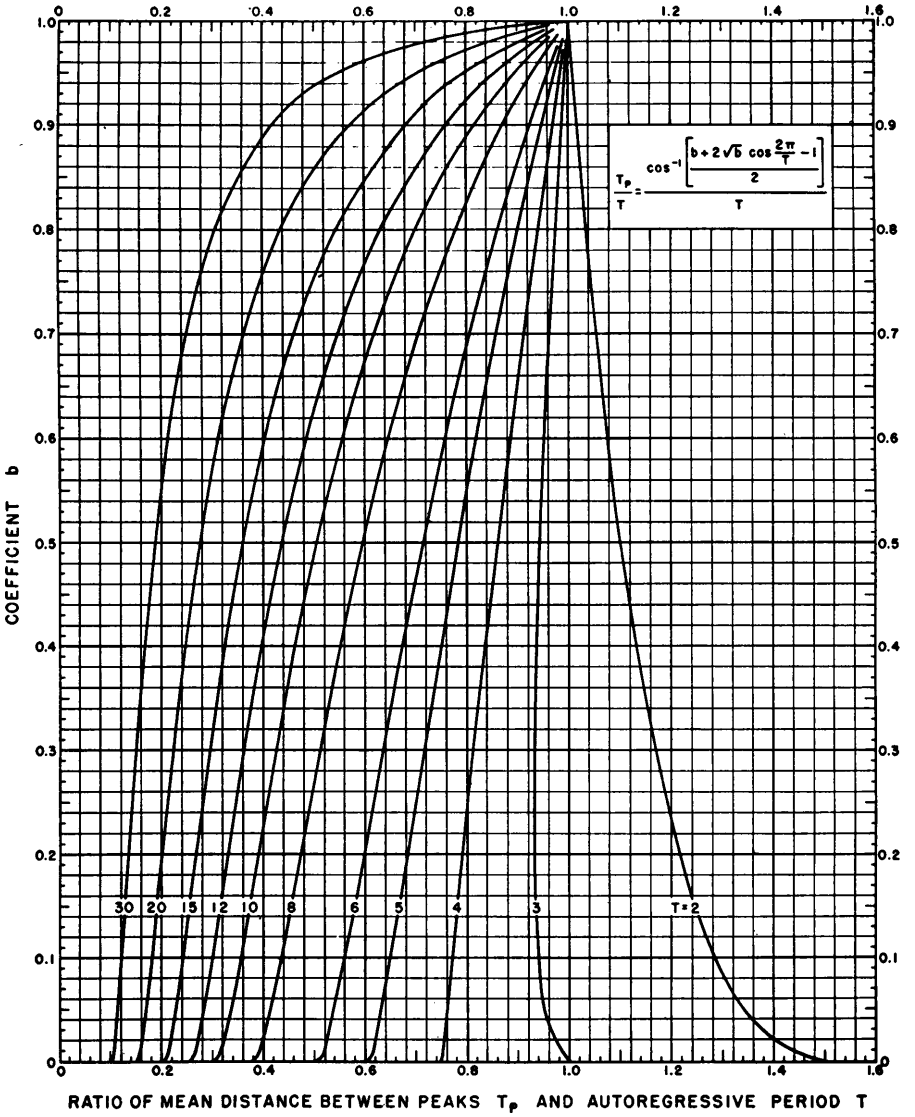

FIGURE 12

Ratio of mean distance between peaks T_p and autoregressive period T

and provide considerably better estimates of the autoregressive period for larger values of the coefficient b.

2. The computed and the observed mean distances between upcrosses were very close to the autoregressive period for values of the coefficient $b = 0.8$ and above. Departures will become greater for increased values of T and for diminishing values of b (figure 13).

3. The agreement between observed and theoretical values is, in general, good; the latter being usually the larger.

The above findings are similar to those from Kendall's [2] experimental study of

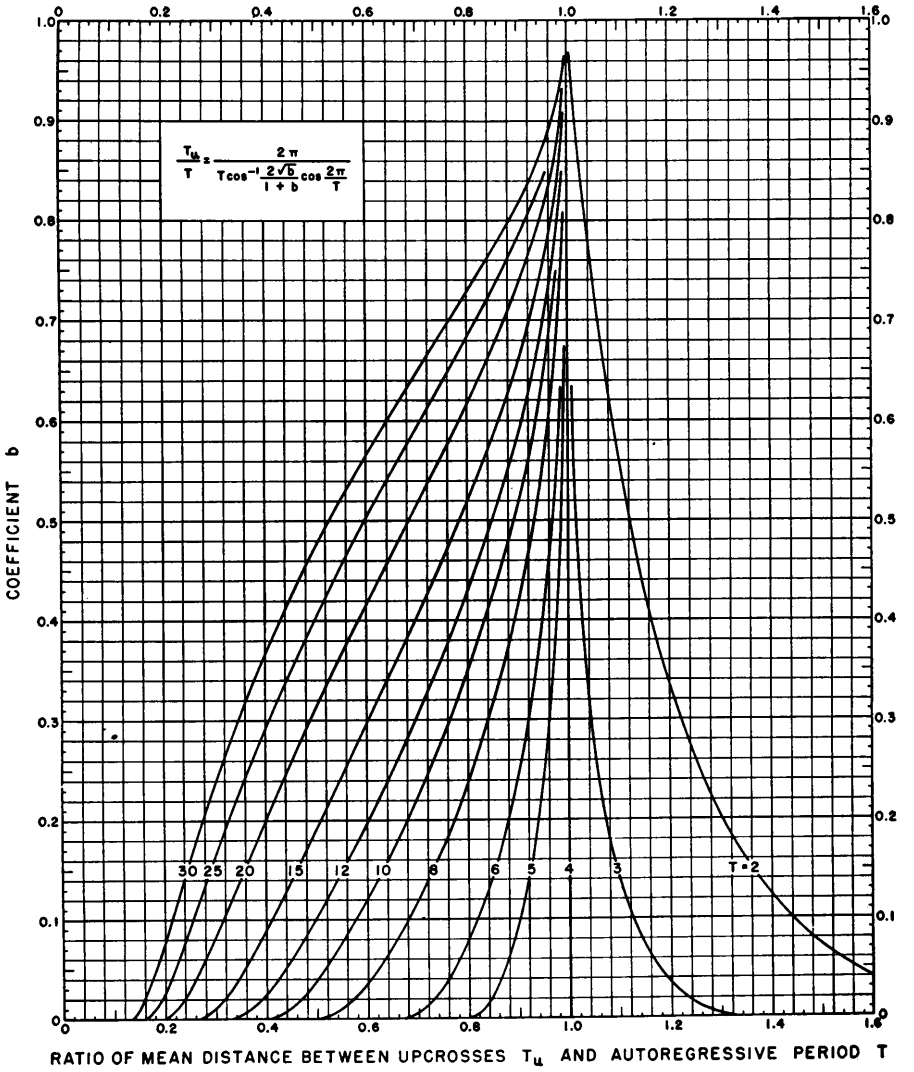

FIGURE 13

Ratio of mean distance between upcrosses T_u and autoregressive period T

four autoregressive series. For completeness, we tabulate certain of his results in table VI showing the ratios of computed peak and upcross distances to the autoregressive periods. With the exception of series 2, there is good agreement with theoretical expectancy (figures 12, 13).

5.4. *Correlogram analyses.* Each of the series was subjected to analysis and the autocorrelation coefficients r_k plotted against the interval, k, form the correlograms

TABLE VI

RATIOS OF COMPUTED PEAK AND UPCROSS DISTANCES TO AUTOREGRESSIVE PERIODS*

Series	Coefficients a	b	Auto Period T	M.D. (Peaks) T_p	T_p/T	M.D. (Upcrosses) T_u	T_u/T
1	−1.2	0.4	9.25	4.96	0.536	8.40	0.908
2	−1.5	0.8	19.53	4.96	0.254	11.61	0.594
3	−1.0	0.6	6.92	5.69	0.822	6.87	0.993
4	−0.8	0.8	3.00	2.60	0.866	2.73	0.910

* Data from Kendall's four experimental autoregressive series.

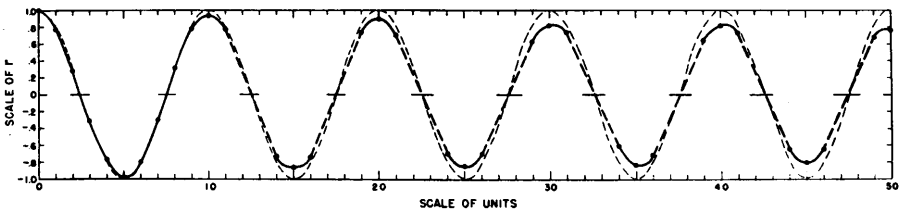

FIGURE 14

Correlogram model III, series B with superimposed theoretical values

of figures 14 to 22. They are referred to as computed correlograms. The theoretical correlograms superimposed on the computed correlograms were obtained from theoretical autocorrelation coefficients computed as follows:

$$r_k = \frac{p^k}{(1+p^2)\sin\theta}\{\sin(k+1)\theta - p^2\sin(k-1)\theta\}. \tag{24}$$

In considering the general properties of the correlograms, figures 14 to 22 reveal that significant differences may exist between the theoretical and those computed from finite amounts of data. However, similarities in key properties of both need be studied for the purpose of inferring, in so far as possible, the physical properties of correlograms of natural time series. In particular, the nature of the damping and the symmetry of the computed correlograms permit certain inferences to be made concerning properties of the basic series.

The theoretical correlogram of an autoregressive series (equation 24) will damp very close to zero within a very few cycles; the rate of damping depending on the magnitude of the coefficient b. In our experiments, theoretical correlograms damped to less than $r = 0.05$ by the end of the fifth cycle for a value of the

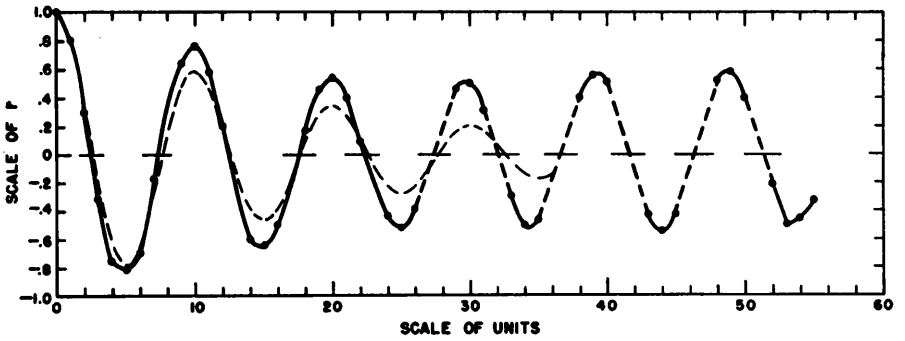

FIGURE 15

Correlogram model III, series F with superimposed theoretical values

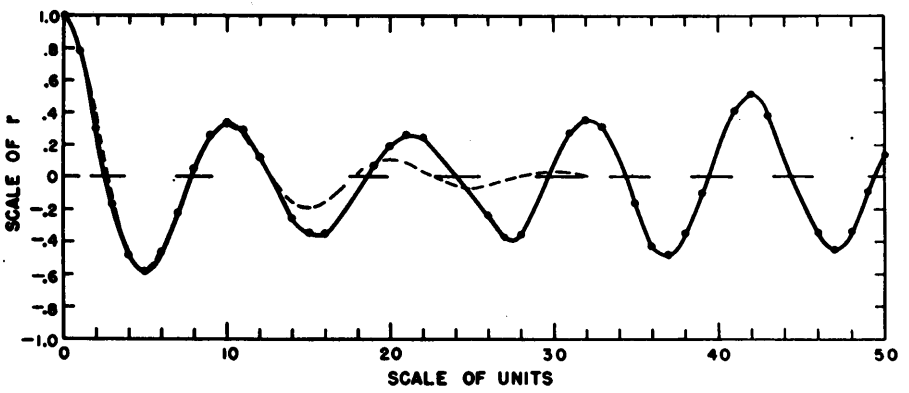

FIGURE 16

Correlogram model III, series G with superimposed theoretical values

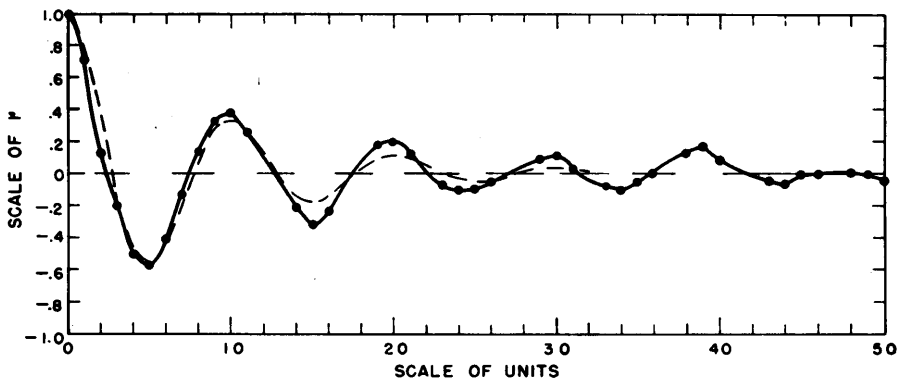

FIGURE 17

Correlogram model III, series G–1 with superimposed theoretical values

coefficient $b = 0.9$ (series F) and by the end of the first cycle when the coefficient b was 0.5 or less.

On the other hand, the computed correlograms of autoregressive series of finite length, may be less strongly damped than expected from theory. This, in general, will depend on the length of the series analyzed and on its damping factor. Our experiments show that when the coefficient b is large, longer series of observations

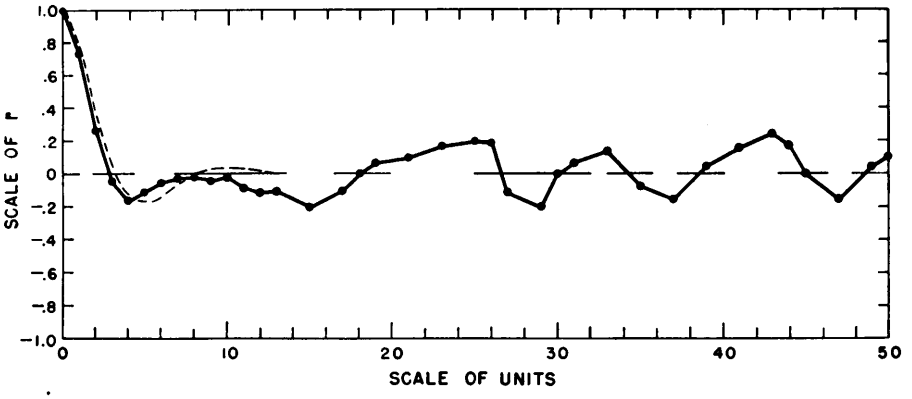

FIGURE 18

Correlogram model III, series I with superimposed theoretical values

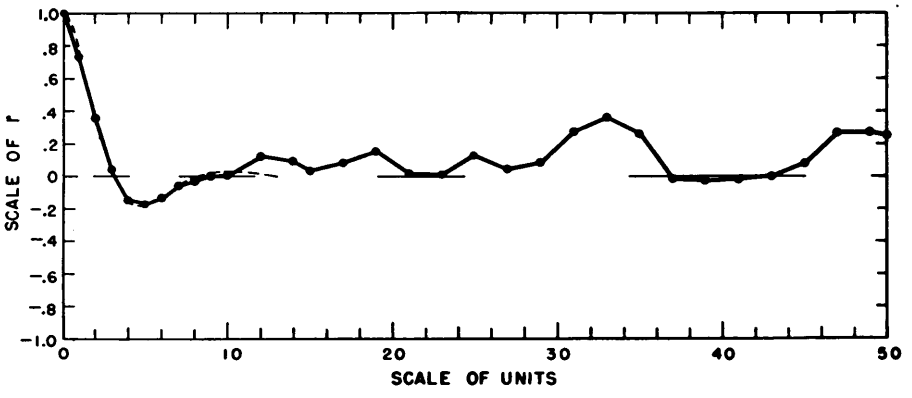

FIGURE 19

Correlogram model III, series J with superimposed theoretical values

are required to approach theoretical damping than when b is small. This is demonstrated (figures 16 and 17) by the series G ($b = 0.80$) correlograms based on 180 and 500 terms, respectively. The correlogram computed from 180 terms damped only to the second cycle and then grew, whereas that of the 500 unit series damped very nearly in accordance with theoretical expectations; at the third cycle, the computed r_k was 0.1 as compared with a theoretical r_k of 0.04. The computed correlogram was completely damped at the end of the fifth cycle.

In the case of the autoregressive series with coefficient $b = 0.5$, or less, correlo-

grams damped to near zero by the end of the first cycle and then grew with irregular oscillatory features. In this, they tend to resemble the theoretical correlograms of mixed sine waves, although the phenomenon results from finiteness of the data.

In series with coefficient $b = 0.6$ and $b = 0.7$, the correlograms based on 400 and 500 units, respectively, damped very nearly in accordance with theoretical expectations (figures 21 and 22).

It appears reasonable that good approximations to theory are obtained by using

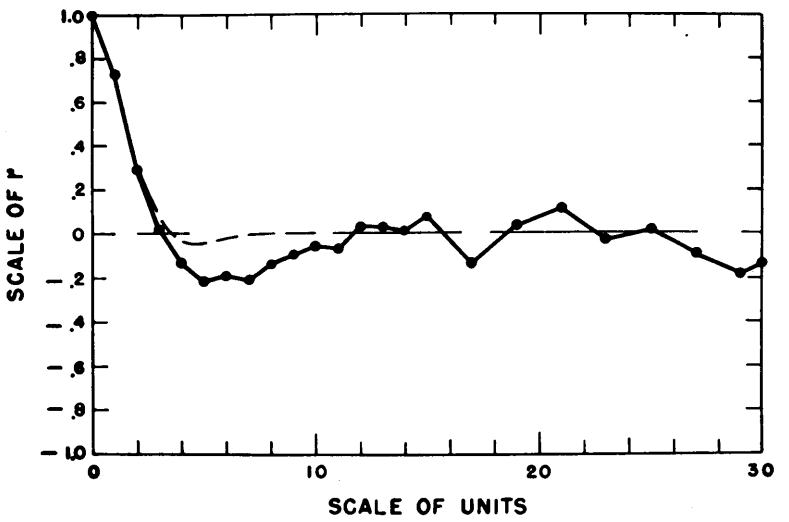

FIGURE 20

Correlogram model III, series K with superimposed theoretical values

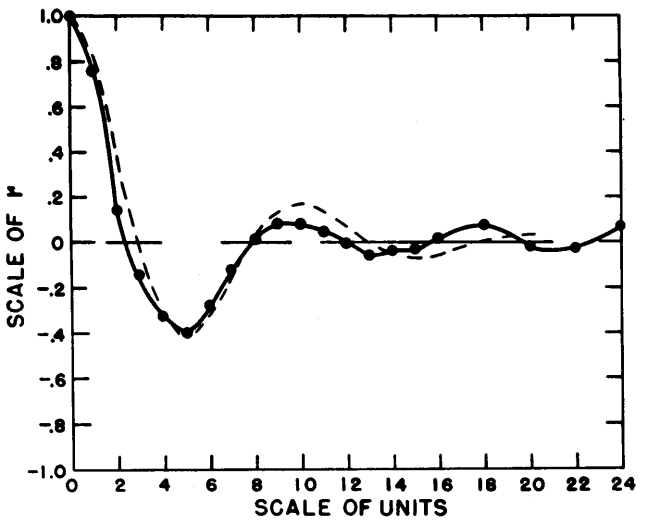

FIGURE 21

Correlogram model III, series L with superimposed theoretical values

400 to 500 (40 to 50 cycles) terms in an autoregressive series. However, when data are not available, failure of the computed correlogram to damp according to theory is not grounds for rejection of the theory of autoregression. This confirms the evidence of Kendall, and it is apparent that the correlogram itself must serve as the final guide for estimating properties of unknown natural time series.

Earlier it was shown that the correlogram for series consisting of a single cyclical component and not completely masked by a random component, symmetrically defines the period, and its terminal amplitude damps close to that required by theory. On the other hand, the correlogram of series composed of several cyclical components is not necessarily symmetrical, it frequently shows minor humps or ripples in the first cycle, but the period is not revealed and a Fourier transform of the autocorrelation is required for additional information. The correlogram of the

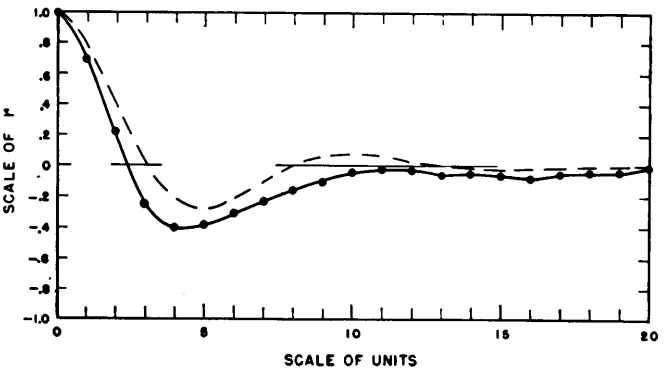

FIGURE 22

Correlogram model III, series M with superimposed theoretical values

autoregressive series is somewhat similar to that of the first case and the autoregressive period usually indicated by the location of the first valley and frequently, also, by its first peak when not too strongly damped.

Estimation of the autoregressive period from the first two computed autocorrelations (equation 11) is so sensitive to departures from theory that in event of significant difference with the period indicated by the first correlogram cycle it should be viewed with suspicion. Tabulation of computed autocorrelation coefficients (table V) shows that at times agreement may be poor, as for instance, for series L and M.

6. Experimental model IV

6.1. *Pattern of analyses.* This section is concerned with an experimental correlogram analysis of five series, each consisting of an autoregressive function to which are added cosine waves of different amplitudes and periods. The autoregressive function (series G of model III) was identical for the first four series (A to D); namely,

$$x_{n+2} = 1.45x_{n+1} - 0.80x_n + \epsilon_{n+2} \tag{25}$$

with ϵ rectangularly distributed between -1 and $+1$. The autoregressive function

of series E was strongly damped (series J of model III). Namely,

$$x'_{n+2} = 1.14x'_{n+1} - 0.50x'_n + \epsilon_{n+2} \tag{26}$$

with ϵ normally distributed between 0 and 1.

The five artificial series analyzed in this section were formed as follows:

Series A: $y = x_{n+2} + \cos \frac{2\pi x}{10}$, 180 units;

Series B: $y = x_{n+2} + 4 \cos \frac{2\pi x}{10}$, 180 units;

Series C: $y = x_{n+2} + 4 \cos \frac{2\pi x}{5}$, 180 units;

Series D: $y = x_{n+2} + 4 \cos \frac{2\pi x}{20}$, 180 units;

Series E: $y = x'_{n+2} + 4 \cos \frac{2\pi x}{10}$, 360 units.

In three of the above cases the period of the cosine was identical with the autoregressive period (10 units). Statistical properties of the series are tabulated in table VII.

TABLE VII

PROPERTIES OF ARTIFICIALLY GENERATED MODEL IV SERIES

Property	Series A	Series B	Series C	Series D	Series E
Mean	−0.01	−0.08	−0.08	−0.08	−0.19
Average deviation	1.24	2.65	2.70	2.73	2.57
Standard deviation	1.53	3.12	3.13	3.19	3.08
Deviation ratio	0.81	0.85	0.86	0.86	0.84
Mean dist. peaks	7.23	9.41	5.00	9.81	7.64
Mean dist. upcrosses	9.93	9.97	5.10	16.03	9.77
Total variance					
computed	2.35	9.73	9.81	10.15	9.50
theoretical	3.19	10.69	10.69	10.69	11.15
Cosine variance					
computed	0.50	8.01	8.01	8.00	8.04
theoretical	0.500	8.00	8.000	8.000	8.00
Autoregressive variance					
computed	1.89	1.89	1.89	1.89	2.84
theoretical	2.68	2.68	2.68	2.68	3.16
r_1	0.79	0.79	0.27	0.85	0.71
r_2	0.35	0.27	−0.54	0.68	0.19

6.2. *Properties of the basic data.* The mean distances between peaks and upcrosses for the autoregressive series G and J of model III are:

		Computed	Theoretical
Series G	M.D. peaks	6.56	7.02
	M.D. upcrosses	9.42	9.91
Series J	M.D. peaks	4.85	5.05
	M.D. upcrosses	7.20	8.88

The computed and theoretical values agreed within one half unit, except for the M.D. upcrosses of the strongly damped series J. The addition of a cosine to the autoregressive changes the mean distances between peaks and upcrosses (table VII) in a fashion that they approach the value of the cosine period. This is particularly noticeable for the series having a cosine period of 5 (series C) and of 20 (series D) units which masks the ten unit autoregressive period.

A property of pure autoregressive series is that the ratio of average deviation to standard deviation lay between 0.78 and 0.81 or very close to $\sqrt{2/\pi}$, a ratio usually characterizing unimodal curves approaching symmetry. The addition of a

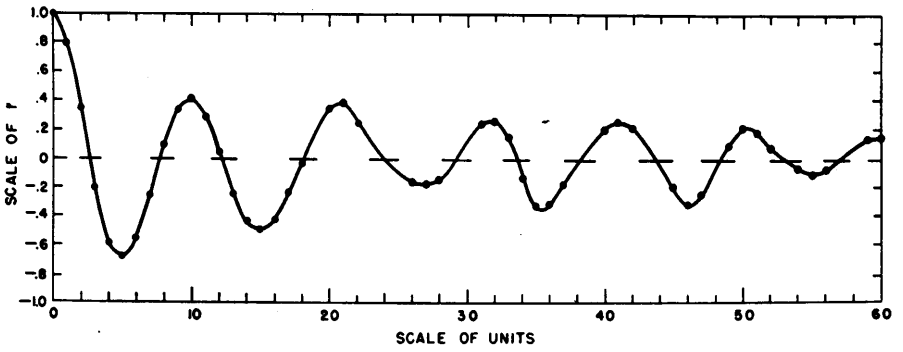

FIGURE 23

Correlogram model IV, series A

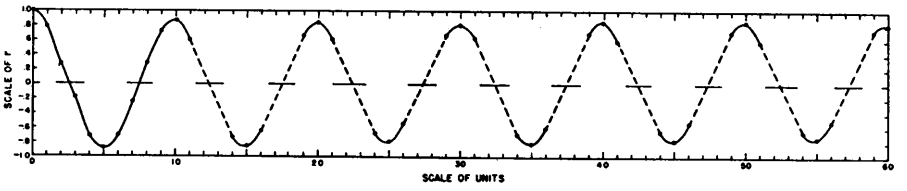

FIGURE 24

Correlogram model IV, series B

cosine to the autoregressive increased this ratio, 0.81 to 0.86, to approximate the AD/σ ratio of $2\sqrt{2}/\pi$ for cosines.

6.3. *Correlogram analyses.* Each of the five series was subjected to analysis as in previous models and the autocorrelation coefficients r_k plotted against the interval k to form the correlograms of figures 23 to 27. We briefly consider their properties and variations from those of pure autoregressive series.

A prominent correlogram feature is the regulation of its period by the cosine. Thus, while the correlogram for the pure autoregressive series G (figure 16) indicated a period of 10 units only for the first cycle (after which, distances from peak to peak and valley to valley became irregular) the addition of a 10 unit cosine strengthened its symmetry. Again, the series E correlogram had uniformly spaced (10 units) peaks and valleys, though more than six cycles, while the series C

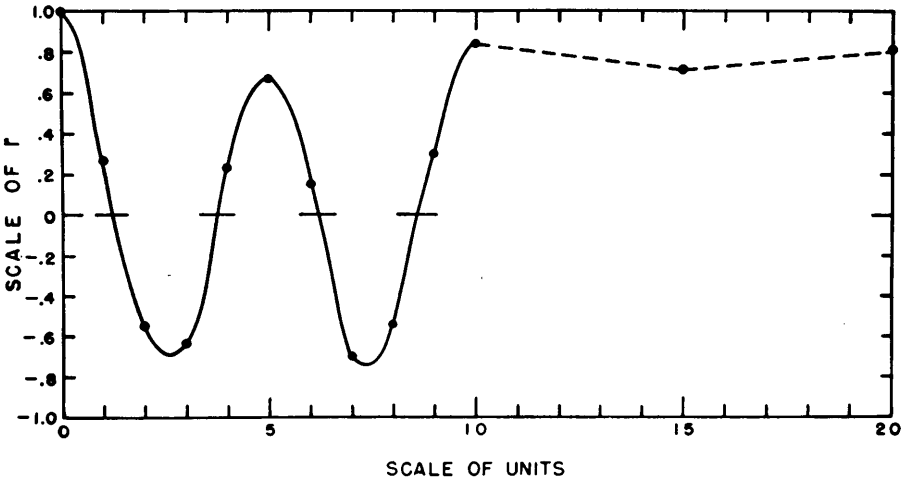

FIGURE 25

Correlogram model IV, series C

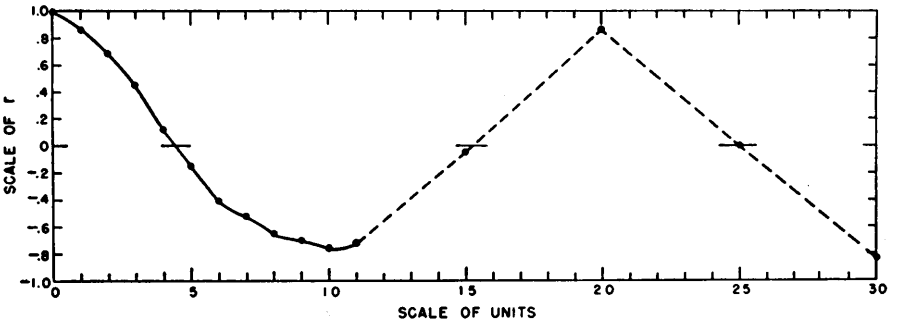

FIGURE 26

Correlogram model IV, series D

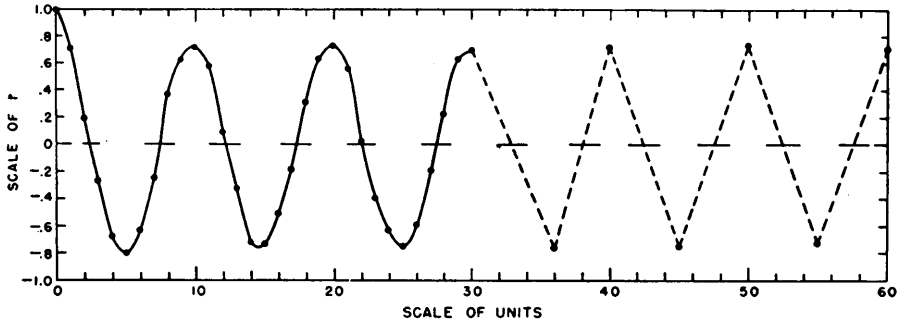

FIGURE 27

Correlogram model IV, series E

correlogram is striking with equally spaced five unit peaks and valleys. Series D revealed its twenty unit cosine period in the first correlogram cycle.

The correlograms of this model will not damp to zero except in the case where a trigonometric should be completely masked by the autoregressive. In this respect, they are similar to those for a sum of sines (model A) and the autoregressive appears to exert the influence of a single trigonometric of like period. The magnitude of the influence is related to its coefficient b and its autoregressive period in relation to the period and amplitude of the trigonometric.

The effect of the added cosine on damping of the correlogram is illustrated from autocorrelation values at the first valley (half cycle, $r_{k/2}$) and the first peak (first cycle, r_k) tabulated in table VIII. Both series B and E correlograms are similar to

TABLE VIII

Model	r_T (th)	r_T (comp)	r_T (obs)	$r_{k/2}$	r_k
III-G	....	....	.....	−0.58	0.33
IV-A	0.16	0.20	(0.15)	−0.67	0.41
B	0.75	0.81	(0.8)	−0.88	0.87
C	0.75	0.81	.....	−0.70	0.66
D	0.75	0.82	.....	−0.76	0 87
III-J	....	....	.....	−0.17	0.13
IV-E	0.72	0.72	(0.7)	−0.80	0.72

those for the combination of a single trigonometric and a random component and damp to terminal amplitudes at the first cycle; series B to about 0.85 and series E to 0.7. On the other hand, series A (smaller amplitude) correlogram damps continuously to the sixth peak when it attains a terminal amplitude of about 0.15. The series C (cosine one half the period of the autoregressive) correlogram shows alternate damping and growth, similar to a combination of two cosines. A similar situation is indicated for the series D correlogram (cosine period twice that of the autoregressive).

REFERENCES

[1] G. E. R. DEACON, "Waves and swell," *Quarterly Jour. Roy. Meteorological Soc.*, Vol. 75, No. 325 (1949), pp. 227–238.

[2] M. G. KENDALL, "Contributions to the study of oscillatory time series," *Occasional Papers IX, National Institute of Economic and Social Research*, Cambridge University Press, Cambridge, 1946.

[3] M. G. KENDALL, *The Advanced Theory of Statistics*, Vol. 2, Griffin, London, 1948.

[4] A. A. KLEBBA, "Details of shore-based wave recorder and ocean wave analyzer," *Annals New York Acad. Sci.*, Vol. 51 (1949), pp. 533–544.

[5] H. R. SEIWELL and G. P. WADSWORTH, "A new development in ocean wave research," *Science*, Vol. 109 (1949), pp. 271–274.

[6] H. R. SEIWELL, "The principles of time series analyses applied to ocean wave data," *Proc. Nat. Acad. Sci.*, Vol. 35, No. 9 (1949), pp. 518–528.

[7] ———, "A new mechanical autocorrelator," *Review of Scientific Instruments*, Vol. 21, No. 5 (1950), pp. 481–484.

[8] E. B. WILSON, "The periodogram of American business activity," *Quarterly Jour. of Economics* (1934), pp. 375–417.

[PRINTED IN U.S.A.]